

Mostly Harmless Econometrics

A Course Correction in the History of Econometrics

Question → Specification & Estimand → Identification → Estimation → Inference

Kaiwen Luo

Washington University in St.Louis

October 16, 2026

A familiar way to learn econometrics

Many traditional econometrics courses are organized around estimators:

OLS $\rightarrow$ GLS $\rightarrow$ MLE $\rightarrow$ IV/2SLS $\rightarrow$ Panel FE $\rightarrow$ Discrete Choice.

A familiar way to learn econometrics

Many traditional econometrics courses are organized around estimators:

OLS $\rightarrow$ GLS $\rightarrow$ MLE $\rightarrow$ IV/2SLS $\rightarrow$ Panel FE $\rightarrow$ Discrete Choice.

This training is valuable. It teaches us how to solve problems such as

$$\min_{\beta} \sum_i (Y_i - X_i' \beta)^2,$$

how to derive

$$\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{d} N(0, V),$$

and how to construct standard errors.

A familiar way to learn econometrics

Many traditional econometrics courses are organized around estimators:

OLS $\rightarrow$ GLS $\rightarrow$ MLE $\rightarrow$ IV/2SLS $\rightarrow$ Panel FE $\rightarrow$ Discrete Choice.

This training is valuable. It teaches us how to solve problems such as

$$\min_{\beta} \sum_i (Y_i - X_i' \beta)^2,$$

how to derive

$$\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{d} N(0, V),$$

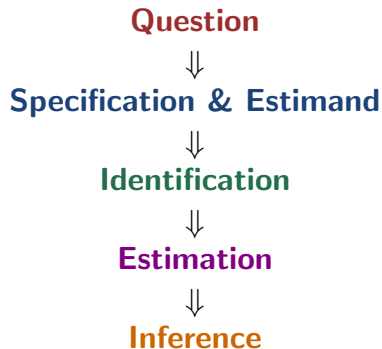
and how to construct standard errors.

But there is a more fundamental question

Before asking how to estimate β , we should ask:

Why is β the object we want in the first place?

The econometric workflow



The central principle of this course

An estimator is not a research design.

A regression equation is not automatically a causal model.

A model that can be estimated is not necessarily a model that answers the question we care about.

What does each step ask?

Question	What economic or causal question are we trying to answer?
Specification	What are the units, outcomes, choices, treatments, states, constraints, and behavioral relationships?
Estimand	What precise population object represents the answer to our question?
Identification	Under what assumptions is that object uniquely determined by the population distribution of observed data?
Estimation	Given a finite sample, how do we construct an estimator of that object?
Inference	How uncertain is our estimate because we observe only a finite sample?

Do not reverse the order

“I can estimate this regression” $\nRightarrow$ “this regression answers my question.”

Specification and identification are different

Suppose a model is indexed by a parameter θ and implies an observable-data distribution P_θ .
A classical definition of identification is:

$$P_{\theta_1} = P_{\theta_2} \implies \theta_1 = \theta_2.$$

Specification and identification are different

Suppose a model is indexed by a parameter θ and implies an observable-data distribution P_θ . A classical definition of identification is:

$$P_{\theta_1} = P_{\theta_2} \implies \theta_1 = \theta_2.$$

Specification comes first:

Which family $\{P_\theta : \theta \in \Theta\}$ are we willing to consider?

“All models are wrong, but some are useful.” — George E. P. Box

Specification and identification are different

Suppose a model is indexed by a parameter θ and implies an observable-data distribution P_θ . A classical definition of identification is:

$$P_{\theta_1} = P_{\theta_2} \implies \theta_1 = \theta_2.$$

Specification comes first:

Which family $\{P_\theta : \theta \in \Theta\}$ are we willing to consider?

“All models are wrong, but some are useful.” — George E. P. Box

Then **identification** asks:

Within that model, can different values of θ generate the same observable data?

Important

Identification can be perfect *within a badly chosen model*.

That does not rescue the specification.

Two examples for today

We will study the same workflow through two very different traditions.

Example 1: Treatment effects

Question: Does vaccination reduce COVID infection?

Language:

$Y(1), Y(0), ATE$

Main flavor:

design-based / reduced-form

Example 2: Binary choice

Question: How will consumers respond to a price change?

Language:

$U_{ij}, \theta, \varepsilon_{ij}$

Main flavor:

structural

The methods look very different.

Two examples for today

We will study the same workflow through two very different traditions.

Example 1: Treatment effects

Question: Does vaccination reduce COVID infection?

Language:

$Y(1), Y(0), ATE$

Main flavor:

design-based / reduced-form

Example 2: Binary choice

Question: How will consumers respond to a price change?

Language:

$U_{ij}, \theta, \varepsilon_{ij}$

Main flavor:

structural

The methods look very different.

But the discipline of reasoning should be the same.

Example I: Start from the question

Suppose someone asks:

“What is the effect of vaccination on COVID infection?”

This sounds like a well-defined causal question.

Example I: Start from the question

Suppose someone asks:

“What is the effect of vaccination on COVID infection?”

This sounds like a well-defined causal question.

It is not.

We still need to specify:

Who?
Which vaccine?
Vaccinated when?
Outcome measured when?
Infection or severe disease?
One person's vaccination or a vaccination campaign?

Lesson

Natural-language questions are usually not yet econometric estimands.

A first specification: individual treatment

Let $D_i \in \{0, 1\}$ indicate whether individual i is vaccinated, and define the potential outcomes

$Y_i(1)$ = infection outcome if i is vaccinated,

$Y_i(0)$ = infection outcome if i is not vaccinated.

Then an apparently natural estimand is

$$\text{ATE} = \mathbb{E}[Y_i(1) - Y_i(0)].$$

Now we have translated a verbal question into a mathematical object.

A first specification: individual treatment

Let $D_i \in \{0, 1\}$ indicate whether individual i is vaccinated, and define the potential outcomes

$Y_i(1)$ = infection outcome if i is vaccinated,

$Y_i(0)$ = infection outcome if i is not vaccinated.

Then an apparently natural estimand is

$$\text{ATE} = \mathbb{E}[Y_i(1) - Y_i(0)].$$

Now we have translated a verbal question into a mathematical object.

But have we translated the question correctly?

That is a **specification question**, not an estimation question.

What is hidden inside $Y_i(1)$ and $Y_i(0)$?

Writing $Y_i(d)$ implicitly assumes that once we specify i 's treatment d , the potential outcome is well defined.

A standard version of SUTVA requires, among other things:

- ① no relevant hidden versions of treatment;
- ② no interference across units.

The second condition means:

Y_i depends on D_i , but not on D_j for $j \neq i$.

Equivalently,

$$Y_i(\mathbf{D}) = Y_i(D_i).$$

What is hidden inside $Y_i(1)$ and $Y_i(0)$?

Writing $Y_i(d)$ implicitly assumes that once we specify i 's treatment d , the potential outcome is well defined.

A standard version of SUTVA requires, among other things:

- ① no relevant hidden versions of treatment;
- ② no interference across units.

The second condition means:

Y_i depends on D_i , but not on D_j for $j \neq i$.

Equivalently,

$$Y_i(\mathbf{D}) = Y_i(D_i).$$

This is not merely technical notation

It is an economic and scientific statement about how the world works.

Vaccination exposes the specification problem

For an infectious disease, individual i 's infection risk can depend on D_i , and also on the vaccination choices of the people around i . For example,

$$Y_i = Y_i(D_i, D_{-i}).$$

Or, using a local vaccination rate $G_i = \frac{1}{|N_i|} \sum_{j \in N_i} D_j$, we may instead need

$$Y_i(d, g).$$

If vaccination changes transmission, then

$$Y_i(0, g = 0.8) \neq Y_i(0, g = 0.2)$$

may be perfectly possible.

Therefore

The simple pair $\{Y_i(1), Y_i(0)\}$ may be too small a model for the substantive question.

The research question determines the estimand

Once interference is allowed, several distinct questions appear.

Direct effect at exposure level g :

$$DE(g) = \mathbb{E}[Y_i(1, g) - Y_i(0, g)].$$

Spillover effect for an untreated person:

$$SE(g_1, g_0) = \mathbb{E}[Y_i(0, g_1) - Y_i(0, g_0)].$$

Policy effect:

$$PE(p_1, p_0) = \mathbb{E}[Y_i(\mathbf{D}(p_1)) - Y_i(\mathbf{D}(p_0))],$$

where p indexes a population vaccination policy.

These are different causal questions

There is no unique object called “the effect of vaccination.”

A perfectly executed experiment can answer the wrong question

Suppose we randomly vaccinate half of the individuals:

$$D_i \perp \{Y_i(1), Y_i(0)\}.$$

Under the simple no-interference model,

$$\mathbb{E}[Y_i \mid D_i = 1] - \mathbb{E}[Y_i \mid D_i = 0] = \text{ATE}.$$

Excellent identification.

A perfectly executed experiment can answer the wrong question

Suppose we randomly vaccinate half of the individuals:

$$D_i \perp \{Y_i(1), Y_i(0)\}.$$

Under the simple no-interference model,

$$\mathbb{E}[Y_i \mid D_i = 1] - \mathbb{E}[Y_i \mid D_i = 0] = \text{ATE}.$$

Excellent identification.

But suppose the policy question is:

What happens when population coverage rises from 20% to 80%?

Then an individually randomized experiment conducted around 50% coverage need not identify that policy effect.

The deeper lesson

Strong identification + wrong estimand = a precise answer to the wrong question.

Only now do we discuss identification

Return temporarily to the simple potential-outcome model:

$$Y_i(1), \quad Y_i(0).$$

Observed outcome:

$$Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0).$$

We want

$$\text{ATE} = \mathbb{E}[Y_i(1) - Y_i(0)].$$

But for every individual we observe only one potential outcome.

This is the fundamental missing-data problem:

	$Y_i(1)$	$Y_i(0)$
$D_i = 1$	✓	?
$D_i = 0$	?	✓

Identification asks:

What assumption allows observed outcomes to reconstruct the missing counterfactual?

Random assignment is an identification strategy

Suppose

$$D_i \perp (Y_i(1), Y_i(0)).$$

Then

$$\mathbb{E}[Y_i(1)] = \mathbb{E}[Y_i \mid D_i = 1], \quad \mathbb{E}[Y_i(0)] = \mathbb{E}[Y_i \mid D_i = 0],$$

and therefore

$$\boxed{\text{ATE} = \mathbb{E}[Y_i \mid D_i = 1] - \mathbb{E}[Y_i \mid D_i = 0].}$$

Notice what happened: the causal estimand

$$\underbrace{\mathbb{E}[Y_i(1) - Y_i(0)]}_{\text{causal estimand}}$$

became equal, under an identification assumption, to an observable quantity

$$\underbrace{\mathbb{E}[Y_i \mid D_i = 1] - \mathbb{E}[Y_i \mid D_i = 0]}_{\text{observable population quantity}}.$$

Identification is not estimation

Once identification is established,

$$ATE = \mathbb{E}[Y \mid D = 1] - \mathbb{E}[Y \mid D = 0],$$

we can estimate it using

$$\widehat{ATE} = \bar{Y}_1 - \bar{Y}_0.$$

This is **estimation**. We can then derive

$$\text{Var}(\widehat{ATE}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_0^2}{n_0},$$

and construct a confidence interval. This is **inference**.

Three logically different statements

Identification: $ATE = \mathbb{E}[Y \mid D = 1] - \mathbb{E}[Y \mid D = 0].$

Estimation: $\widehat{ATE} = \bar{Y}_1 - \bar{Y}_0.$

Inference: $\widehat{ATE} \pm 1.96 SE(\widehat{ATE}).$

Standard errors do not measure identification uncertainty

Suppose the estimate is

$$\widehat{ATE} = -0.08, \quad SE(\widehat{ATE}) = 0.01.$$

This tells us that sampling noise is relatively small.

It does **not** tell us whether

$$D_i \perp (Y_i(1), Y_i(0))$$

is credible.

It does not tell us whether SUTVA holds.

It does not tell us whether we selected the correct outcome horizon.

It does not tell us whether the estimand corresponds to the policy question.

Inference is conditional

Conventional confidence intervals quantify sampling uncertainty *conditional on the maintained model and identification assumptions*.

Example I: the complete workflow

Question	What aspect of vaccination are we trying to evaluate?
Specification	Individual treatment? Population coverage? Network exposure? What outcome and horizon?
Estimand	ATE? Direct effect? Spillover effect? Policy effect?
Identification	Random assignment? No interference? Cluster randomization? Two-stage saturation design?
Estimation	Difference in means? Regression adjustment? IPW?
Inference	Individual-level SE? Clustered SE? Randomization inference?

The estimator is near the end of the argument.

Example II: A structural question

Now consider a different research question.

Suppose consumers must choose between products A and B .

We want to know:

How would demand change if the price of A fell by \$5?

Perhaps we also want:

consumer surplus,

substitution patterns,

or responses to a price never observed in the data.

These questions naturally invite a model of

how choices are generated.

A standard binary-choice specification

Suppose utility from the two alternatives is

$$U_{iA} = X'_{iA}\beta - \alpha p_A + \varepsilon_{iA}, \quad U_{iB} = X'_{iB}\beta - \alpha p_B + \varepsilon_{iB}.$$

Consumer i chooses A if $U_{iA} > U_{iB}$. Define the latent index

$$\Delta U_i = U_{iA} - U_{iB}.$$

Then the observed choice is

$$D_i = \mathbf{1}\{\Delta U_i > 0\}.$$

This is more than a statistical regression.

It specifies a **decision rule** connecting primitives to behavior.

From economic structure to an observable probability

Write

$$\Delta U_i = \Delta X_i' \beta - \alpha \Delta p_i + \nu_i.$$

Then

$$\mathbb{P}(D_i = 1 \mid X_i, p_i) = \mathbb{P}(\nu_i > -\Delta X_i' \beta + \alpha \Delta p_i).$$

If ν_i has the logistic distribution,

$$\mathbb{P}(D_i = 1 \mid X_i, p_i) = \Lambda(\Delta X_i' \beta - \alpha \Delta p_i).$$

If ν_i is normal, we obtain a probit model.

The distributional assumption has economic content

It determines how unobserved utility heterogeneity maps into observed choice probabilities.

A hidden assumption: almost everyone has a strict ranking

Under a standard continuous random-utility specification,

$$\mathbb{P}(U_{iA} = U_{iB}) = 0.$$

Almost every consumer strictly prefers one alternative. Thus $D_i = 1$ is interpreted as evidence that $U_{iA} > U_{iB}$.

But suppose the empirical setting is different:

A substantial fraction of consumers genuinely regard A and B as equally good.

Perhaps the products are almost identical.

Perhaps many consumers explicitly report “no preference.”

Perhaps the observed purchase is determined by a default, position on the screen, or random tie-breaking.

The model can fit the data and still misrepresent the question

A standard logit can easily generate $\mathbb{P}(D_i = A) \approx 0.5$. But it interprets this as arising from variation in ν_i , and therefore from consumers with different *strict* latent preferences.

That is conceptually different from saying:

$$\mathbb{P}(U_{iA} = U_{iB}) > 0.$$

These models may fit current market shares similarly. But they can imply very different predictions for

small price changes, defaults, consumer welfare.

Specification again comes before estimation

MLE cannot tell us whether the behavioral model asks the right question.

An alternative specification with an indifference region

Suppose latent relative utility is

$$V_i = \Delta X_i' \beta - \alpha \Delta p_i + \nu_i.$$

Let $\kappa \geq 0$ denote an indifference threshold.

Define the latent preference state

$$S_i = \begin{cases} A, & V_i > \kappa, \\ I, & |V_i| \leq \kappa, \\ B, & V_i < -\kappa. \end{cases}$$

Now there are three economically distinct states:

prefer A ,	indifferent,	prefer B .
--------------	--------------	--------------

If the consumer is forced to make a binary purchase, we still need a **tie-breaking rule** for state I .

Observed binary choice now contains two mechanisms

Suppose an indifferent consumer chooses A with probability r_i .

Then

$$\mathbb{P}(D_i = A \mid X_i) = \mathbb{P}(V_i > \kappa \mid X_i) + r_i \mathbb{P}(|V_i| \leq \kappa \mid X_i).$$

Observed demand therefore reflects:

preference + indifference + tie-breaking.

If the true world has all three mechanisms but we estimate a standard binary logit, we may attribute all observed behavior to preference.

That changes the interpretation of α and β , and potentially the predicted response to policy.

Identification now becomes a serious question

Suppose all we observe for each consumer is (D_i, X_i, p_i) . Can we separately identify

$$\beta, \quad \alpha, \quad \kappa, \quad r?$$

Not necessarily. Different combinations of

preference heterogeneity, indifference, tie-breaking

may generate the same observed binary choice probabilities.

Additional information may be useful:

randomized prices, repeated choices,

randomized defaults, an explicit “indifferent” option,

additional moments or exclusion restrictions.

Identification is not “the optimizer converged”

Suppose we write a likelihood

$$L(\theta) = \prod_{i=1}^n P_{\theta}(D_i \mid X_i),$$

and obtain $\hat{\theta} = \arg \max_{\theta} L(\theta)$. The computer reports:

Convergence achieved.

Identification is not “the optimizer converged”

Suppose we write a likelihood

$$L(\theta) = \prod_{i=1}^n P_{\theta}(D_i \mid X_i),$$

and obtain $\hat{\theta} = \arg \max_{\theta} L(\theta)$. The computer reports:

Convergence achieved.

That tells us nothing by itself about identification. Identification is a population property:

$$P_{\theta_1}(D \mid X) = P_{\theta_2}(D \mid X) \implies \theta_1 = \theta_2?$$

Optimization and identification are different problems

A unique numerical optimum in one sample does not prove that the economic primitives are identified.

Why estimate a structural binary-choice model?

If the primitives are credibly specified and identified, we can ask counterfactual questions. For example, under a new price $p'_A = p_A - 5$, we can construct

$$\mathbb{P}_{\hat{\theta}}(D_i(p'_A) = A) .$$

More generally, a structural model may allow us to study

policies not directly observed in the historical data.

Potential outputs include

counterfactual demand, consumer surplus, substitution patterns,

and equilibrium outcomes.

The payoff is greater extrapolation. The cost is stronger modeling assumptions.

Example II: the complete workflow

Question	How would demand and welfare respond to a price policy?
Specification	Utility functions, alternatives, information, choice rule, indifference, tie-breaking.
Estimand	Price coefficient? Preference distribution? Elasticity? Counterfactual market share? Welfare?
Identification	What variation separates price sensitivity, heterogeneity, indifference, and tie-breaking?
Estimation	MLE? GMM? Simulated moments? Bayesian methods?
Inference	Standard errors? Bootstrap? Confidence sets for counterfactuals?

A structural model does not escape the same five-step discipline.

Why “reduced form” causes so much confusion

In modern applied economics, people say things like

“This is a reduced-form paper.”

or

“We first show the reduced-form effect.”

But the term has an older and much more precise econometric meaning.

Important distinction

We should distinguish

reduced form as a mathematical object

from

“reduced-form empirical work” as a research style.

The strict econometric definition

Consider a simultaneous structural system

$$A_0 Y_i = A_1 X_i + U_i,$$

where Y_i collects the endogenous variables and X_i the exogenous variables. If A_0 is invertible,

$$Y_i = A_0^{-1} A_1 X_i + A_0^{-1} U_i.$$

Define $\Pi = A_0^{-1} A_1$ and $V_i = A_0^{-1} U_i$. Then

$$\boxed{Y_i = \Pi X_i + V_i}$$

is the **reduced form**.

Every endogenous variable is expressed as a function of exogenous variables and disturbances.

Example: demand and supply

Consider

$$Q^d = \alpha_0 - \alpha_1 P + \alpha_2 Z_d + u_d,$$

$$Q^s = \beta_0 + \beta_1 P + \beta_2 Z_s + u_s.$$

Equilibrium requires

$$Q^d = Q^s = Q.$$

The two equations are **structural**:

one describes demand behavior, the other supply behavior.

Price P and quantity Q are jointly determined endogenous variables.

Solving the structural system produces the reduced form

Set demand equal to supply:

$$\alpha_0 - \alpha_1 P + \alpha_2 Z_d + u_d = \beta_0 + \beta_1 P + \beta_2 Z_s + u_s.$$

Therefore,

$$P = \frac{\alpha_0 - \beta_0 + \alpha_2 Z_d - \beta_2 Z_s + u_d - u_s}{\alpha_1 + \beta_1}.$$

Hence

$$P = \pi_0 + \pi_d Z_d + \pi_s Z_s + v_P.$$

Substituting this expression into either structural equation gives

$$Q = \gamma_0 + \gamma_d Z_d + \gamma_s Z_s + v_Q.$$

These equations are the reduced form.

Structural and reduced-form parameters are different objects

For example,

$$\pi_d = \frac{\alpha_2}{\alpha_1 + \beta_1}.$$

Thus a reduced-form coefficient generally combines several structural parameters.

Observing π_d does not automatically tell us α_1 , the slope of demand.

To recover structural parameters we need additional restrictions, variation, or instruments.

This is the classical identification problem

Can the structural parameters be recovered from the reduced-form distribution?

The word “reduced form” in modern applied micro

Modern applied researchers often use the term more broadly.
Suppose Z_i is experimentally or quasi-experimentally assigned.
They may estimate

$$Y_i = \alpha + \rho Z_i + \varepsilon_i$$

and call $\hat{\rho}$ a **reduced-form effect**.

Why?

Because the researcher is estimating the response of an observed outcome to plausibly exogenous variation without fully modeling the behavioral mechanism connecting Z to Y .

But this is a convention, not the general mathematical definition

OLS is not synonymous with reduced form.

IV makes the terminology especially transparent

Suppose encouragement Z_i affects treatment D_i .

The first stage is

$$D_i = \pi_0 + \pi_1 Z_i + v_i.$$

The outcome reduced form is

$$Y_i = \rho_0 + \rho_1 Z_i + e_i.$$

The Wald estimand is

$$\frac{\rho_1}{\pi_1}.$$

Under the appropriate IV assumptions, this may identify a LATE.

Here “reduced form” is close to the classical meaning:

an endogenous outcome is written directly as a function of exogenous variation.

Four common misconceptions

❶ **“OLS means reduced form.”**

False. OLS is an estimation method.

❷ **“Logit means structural.”**

False. Logit is a probability model. It becomes structural only when its objects have a behavioral interpretation within an economic model.

❸ **“Structural models must be nonlinear or complicated.”**

False. A structural relationship can be linear.

❹ **“Reduced form means atheoretical.”**

False. A credible reduced-form design requires theory about assignment, counterfactuals, institutions, and the estimand.

So how do we tell them apart in practice?

The misconceptions share one lesson: the econometric *tool* does not settle the classification.

First, accept a premise

This is not a binary classification. It is a **spectrum**, and many of the best papers sit somewhere in the middle.

design-based $\longleftrightarrow$ fully structural
many good papers live in between

The useful question is not “which estimator?” but

What role does the economic model play in the paper?

Four diagnostic questions:

- ① Where does the estimating equation come from?
- ② What is the target parameter?
- ③ What does identification rely on?
- ④ What are the estimates used for?

Q1: Where does the estimating equation come from?

Reduced form

Dictated mainly by the **identification strategy**: the DiD double difference, the comparison around an RD cutoff.

There may be a “theoretical framework” section for intuition. Delete it, and the regression still runs and the result can still be interpreted.

Structural

Derived from an **explicit economic model**: objectives, constraints, information, equilibrium conditions.

The estimating equations are its first-order conditions, likelihood, or moment conditions.

A useful test: take the model away

Do the estimates still mean anything?

- Yes $\Rightarrow$ the paper leans reduced form.
- The parameters cannot even be defined without the model $\Rightarrow$ structural.

In our examples: $\bar{Y}_1 - \bar{Y}_0$ survives the test; α as a “marginal utility of money” does not.

Q2: What is the target parameter?

Reduced form

Treatment effects, elasticities: “in *this* setting, when X changes, how does Y change?”

A composite of many primitives *and* the environment.

Structural

Primitives: preferences, technology, costs, discount rates, risk aversion, beliefs.

In principle, *invariant to the policy* being evaluated.

Recall the demand–supply example:

$$\pi_d = \frac{\alpha_2}{\alpha_1 + \beta_1}, \quad \begin{array}{l} \alpha_1, \alpha_2 : \text{demand primitives,} \\ \beta_1 : \text{supply primitive.} \end{array}$$

π_d is a legitimate causal effect of Z_d on P *in this market*. A policy that changes the supply slope β_1 changes π_d , but leaves α_1, α_2 untouched.

Why this matters

Reduced-form effects describe this environment; primitives can be carried to a new one.

Q3: What does identification rely on?

Reduced form

Mainly the **research design**: exogenous variation in the data (randomization, discontinuities, timing of reforms).

Structural

To a large extent the **model itself**: functional forms, distributional assumptions, equilibrium conditions, exclusion restrictions.

Caveat. Good structural papers *also* use quasi-experimental variation. So the question is not “design or model?” but

How much of the identification burden does the model carry?

In our examples:

- Example I: randomization does essentially all the work.
- Example II: with only (D_i, X_i, p_i) , separating α, β, κ, r leans heavily on the model, unless we add randomized prices, randomized defaults, or an explicit “indifferent” option.

Q4: What are the estimates used for?

This is often the most visible signal: look at where the paper *ends*.

Reduced form

The endpoint is the **effect itself**: report it and interpret it.

Structural

The endpoint is **counterfactual simulation**:

- policies never observed in the data;
- welfare analysis;
- decomposing mechanisms.

In Example II the payoff was

$$\mathbb{P}_{\hat{\theta}}(D_i(p'_A) = A), \quad p'_A = p_A - 5,$$

a price that may never appear in the data.

Use the four questions together

Ask all four about any paper and you can place it on the spectrum, whatever estimator it happens to use.

Summary: the two ends of the spectrum

	Reduced-form / design-based	Structural
Estimating equation	Dictated by the research design	Derived from an explicit economic model
Target parameter	Treatment effects, elasticities in this setting	Primitives: preferences, technology, costs, beliefs
Identification	Mainly exogenous variation in the data	Model structure + identifying variation
End use	Report and interpret the effect	Counterfactuals, welfare, mechanisms
Strength	Transparent connection between variation and estimand	Mechanisms and richer counterfactuals
Main risk	External validity / limited counterfactual scope	Misspecification / stronger extrapolation assumptions

Real papers mix the two columns. The question is *how much work the model does*.

Same empirical setting, different questions

Suppose a city randomly offers a \$5 public-transit subsidy.

Reduced-form question: the effect

$$\mathbb{E}[Y_i(1) - Y_i(0)],$$

where treatment means receiving the subsidy offer. We might ask:

How much did ridership change because of this program?

Structural question: estimate preferences over travel time, money, comfort, and modes. Then ask:

What would happen under a \$2, \$10, or nonlinear subsidy, or after adding a new transit mode?

Same market. Different estimands. Different assumptions.

Reduced form versus structural is not a contest

It is tempting to ask:

Which approach is better?

That is usually the wrong question.

The correct question is:

What do we want to learn, and what assumptions are required to learn it?

Structural work can use randomized experiments.

Reduced-form evidence can discipline structural models.

Experimental moments can identify structural parameters.

Structural models can clarify what a reduced-form coefficient means.

Modern empirical work increasingly combines the two

The boundary is methodological, not ideological.

Two stories about the same regression

Every course writes down the same equation:

$$Y_i = X_i' \beta + \varepsilon_i.$$

But different courses tell different stories about *where the randomness comes from*.

Story 1: fixed design

“ X is fixed in repeated samples.”

$$Y_i = x_i' \beta + \varepsilon_i, \quad x_i \text{ nonrandom.}$$

Only ε_i is random.

Story 2: random design

“ (Y_i, X_i) is an i.i.d. draw from a population.”

$$(Y_i, X_i) \sim F, \quad X_i \text{ random.}$$

Both are random.

And yet nearly every formula carries a $| X$

$$\mathbb{E}[\hat{\beta} | \mathbf{X}], \quad \text{Var}(\hat{\beta} | \mathbf{X}), \quad \mathbb{E}[\varepsilon_i | X_i] = 0.$$

Why do we condition on something we just called fixed?

Fixed design: the experimenter chooses $\mathbf{X}$

The fixed-design convention was inherited from experimental statistics. In an agricultural field trial the researcher *chooses* the fertilizer doses

$$x_1, \dots, x_n,$$

assigns them to plots, and observes yields. Here “repeated samples” is literal: run the experiment again *at the same $\mathbf{X}$* .

Then there is nothing to condition on, and the classical results are exact:

$$\mathbb{E}[\hat{\beta}] = \beta, \quad \text{Var}(\hat{\beta}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.$$

This is the world in which the Gauss–Markov theorem is usually stated.

Fixed design: the experimenter chooses $\mathbf{X}$

The fixed-design convention was inherited from experimental statistics. In an agricultural field trial the researcher *chooses* the fertilizer doses

$$x_1, \dots, x_n,$$

assigns them to plots, and observes yields. Here “repeated samples” is literal: run the experiment again *at the same $\mathbf{X}$* .

Then there is nothing to condition on, and the classical results are exact:

$$\mathbb{E}[\hat{\beta}] = \beta, \quad \text{Var}(\hat{\beta}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.$$

This is the world in which the Gauss–Markov theorem is usually stated.

The difficulty for economists

We almost never choose $\mathbf{X}$.

Nobody assigns years of schooling, firm size, or a country's institutions.

Random design: what actually changes

Under random sampling, “another sample” means **drawing new units**, whose regressor values would themselves be different.

So

$$\mathbf{X}'\mathbf{X}, \quad (\mathbf{X}'\mathbf{X})^{-1}, \quad \hat{\beta}$$

are all random, and in general

$$\mathbb{E} [(\mathbf{X}'\mathbf{X})^{-1}] \neq (\mathbb{E}[\mathbf{X}'\mathbf{X}])^{-1}.$$

Two questions now come apart:

Conditional: how would $\hat{\beta}$ vary across samples with *this* $\mathbf{X}$?

Unconditional: how would $\hat{\beta}$ vary across samples with a *new* $\mathbf{X}$?

These are different reference sets

They are two different thought experiments, and they have two different answers.

Conditioning on $\mathbf{X}$ restores the fixed-design algebra

Write

$$\hat{\beta} = \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon.$$

Conditional on $\mathbf{X}$ this is linear in ε , exactly as in the fixed design. So if

$$\mathbb{E}[\varepsilon \mid \mathbf{X}] = 0,$$

then

$$\mathbb{E}[\hat{\beta} \mid \mathbf{X}] = \beta, \quad \text{Var}(\hat{\beta} \mid \mathbf{X}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \quad \text{under } \mathbb{E}[\varepsilon\varepsilon' \mid \mathbf{X}] = \sigma^2 I,$$

and by the law of iterated expectations,

$$\mathbb{E}[\hat{\beta}] = \mathbb{E}[\mathbb{E}[\hat{\beta} \mid \mathbf{X}]] = \beta.$$

How to read the textbook sentence

“Treat X as fixed” is an instruction to **condition on X** .

It is not a claim that X is physically nonrandom.

Is conditioning on X legitimate?

Under random sampling the joint density factors:

$$f(y_i, x_i; \beta, \gamma) = \underbrace{f(y_i | x_i; \beta)}_{\text{contains } \beta} \cdot \underbrace{f(x_i; \gamma)}_{\text{does not}}.$$

If β and γ are variation free, the marginal distribution of X carries no information about β : X is **ancillary**. The conditionality principle then says inference about β should be carried out in the conditional model, i.e. treating the realized $\mathbf{X}$ as the reference set.

When this argument fails

- X is determined jointly with Y (simultaneity): then $f(x)$ depends on β .
- Sampling depends on Y (choice-based or endogenous stratified sampling).
- The target population differs from the sampled one, so the estimand averages over a *different* distribution of X .

A historical note

The fixed-design convention entered econometrics from experimental statistics, where the researcher really does choose X . Economists inherited the algebra without inheriting the experiment.

For a long time “ X is fixed in repeated samples” was taught as an assumption about the world. It is better understood as a choice of reference set:

which repeated experiment are we imagining?

Modern treatments restate the linear model under random sampling, derive results conditionally and then unconditionally, and make heteroskedasticity-robust inference the default. The design-based literature adds a third reference set: randomization itself.

Back to the workflow

“Where does the randomness come from?” belongs to **Specification & Estimand**, long before anyone types `reg y x, robust`.

The same five questions apply to both traditions

	Treatment-effect example	Binary-choice example
Question	Does vaccination change infection risk?	How does price affect choice and welfare?
Specification	Potential outcomes, treatment, interference, horizon	Utility, information, choice set, tie-breaking
Estimand	ATE, direct effect, spillover, policy effect	Preferences, elasticity, welfare, counterfactual share
Identification	Randomization / conditional independence / design	Exclusion restrictions, price variation, functional restrictions
Estimation	Difference in means, regression, weighting	MLE, GMM, simulated moments
Inference	SEs, clustering, randomization inference	Asymptotic SEs, bootstrap, counterfactual uncertainty

Two different ways a project can fail

Failure 1: Wrong identification

The estimand is meaningful, but the data do not identify it.

For example, ATE is the right question, but treatment selection is endogenous.

Problem:

counterfactual is not identified

Failure 2: Wrong specification

The parameter is perfectly identified, but it is not the object corresponding to the substantive question.

For example, estimate $\mathbb{E}[Y(1) - Y(0)]$ when the policy operates through network spillovers.

Problem:

we precisely answer the wrong question

A useful hierarchy of econometric questions

When reading a paper, do not begin with:

“What regression did they run?”

Instead ask:

- ① What is the substantive question?
- ② What mathematical object represents the answer?
- ③ What assumptions make that object meaningful?
- ④ What assumptions connect it to observed data?
- ⑤ What variation actually identifies it?
- ⑥ What estimator implements that identification argument?
- ⑦ What uncertainty does the reported standard error capture?
- ⑧ What important uncertainty does it *not* capture?

When you see a regression, ask what role it plays

Consider

$$Y_i = \alpha + \beta D_i + X_i' \gamma + \varepsilon_i.$$

The same equation can play very different roles. It might be

a descriptive conditional expectation,

an estimator implementing a randomized experiment,

a selection-on-observables identification strategy,

one equation inside a structural model.

The syntax of the regression does not reveal its scientific meaning

The meaning comes from the specification and identification argument.

The regression coefficient is not the starting point

A common workflow is:

write regression $\rightarrow$ estimate $\beta \rightarrow$ interpret β .

A better workflow is:

Question $\rightarrow$ Estimand $\rightarrow$ Identification

and only then

choose a statistical procedure that estimates the identified object.

Modern developments in

IV, DiD, RDD, treatment effects, structural estimation
are much easier to understand from this perspective.

A warning about “assumptions”

Econometric assumptions play different roles.

Specification assumptions:

$$Y_i(d), \quad U_{ij} = X'_{ij}\beta + \varepsilon_{ij}, \quad \text{choice set,} \quad \text{equilibrium concept.}$$

These determine what world we are modeling.

Identification assumptions: $D \perp (Y(1), Y(0))$, exclusion restrictions, rank conditions, monotonicity, support conditions, etc.

These determine what can be learned from observables.

Inference assumptions:

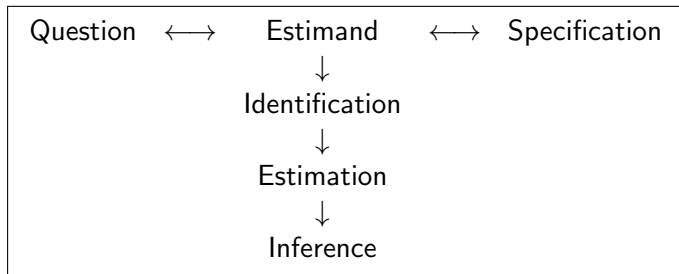
sampling structure, clustering, dependence, regularity conditions.

Do not put all assumptions into one undifferentiated list

They answer different logical questions.

What does “credible” mean?

A credible empirical result requires alignment across all stages:



A sophisticated estimator cannot compensate for
a poorly defined question, an inappropriate model,
or
an unconvincing counterfactual.

The main methodological habit

Before touching the data, try to complete the following sentence:

*"I want to learn _____,
and I can learn it from these data because _____."*

The first blank is primarily about

Question + Specification + Estimand.

The second blank is primarily about

Identification.

Only afterward should we ask:

How should I estimate it?

A roadmap for the rest of the course

We will repeatedly use the same framework.

Question → Specification & Estimand → Identification → Estimation → Inference

Examples will include:

Regression and selection on observables

Instrumental variables and LATE

Difference-in-differences

Regression discontinuity

Panel data

Treatment-effect heterogeneity

Structural models and policy counterfactuals

The methods will change.

The sequence of questions should not.

Takeaways

- ① Econometrics begins with a **question**, not an estimator.
- ② An **estimand** is the precise mathematical object that answers the question.
- ③ **Specification** determines what world we are modeling.
- ④ **Identification** determines what can be learned from the population distribution of observables.
- ⑤ **Estimation** and **inference** come afterward.
- ⑥ Reduced-form and structural methods are not rival religions but two ends of a spectrum: classify a paper by the role its model plays, not by its estimator.

Do not ask
“Which estimator should I use?”
until you know
“What exactly am I trying to learn?”

Selected references

- Angrist, J. D., and J.-S. Pischke (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.
- Heckman, J. J. (2010). "Building Bridges between Structural and Program Evaluation Approaches to Evaluating Policy." *Journal of Economic Literature*, 48(2), 356–398.
- Heckman, J. J., and E. Vytlačil (2005). "Structural Equations, Treatment Effects and Econometric Policy Evaluation." *Econometrica*, 73(3), 669–738.
- Low, H., and C. Meghir (2017). "The Use of Structural Models in Econometrics." *Journal of Economic Perspectives*, 31(2), 33–58.
- Hudgens, M. G., and M. E. Halloran (2008). "Toward Causal Inference with Interference." *Journal of the American Statistical Association*, 103(482), 832–842.
- White, H. (1980). "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity." *Econometrica*, 48(4), 817–838.
- Abadie, A., S. Athey, G. W. Imbens, and J. M. Wooldridge (2020). "Sampling-Based versus Design-Based Uncertainty in Regression Analysis." *Econometrica*, 88(1), 265–296.

**We choose to learn econometrics,
not because it is easy,
but because it is hard.**

— after John F. Kennedy, Rice University, 1962